Method and System for Translating Text

FIELD OF THE INVENTION

The invention relates to a method of translating text sentence from one language to a second language, more particularly, the present invention relates to online translation of web pages over the Internet.

BACKGROUND OF THE INVENTION

For purposes of this disclosure, by the term "network" is meant include at least two computers connected through a physical communication line which can be hardwired, or virtual, such as satellite, cellular or other wireless communications. Computer can mean a personal computer, server or other similar-type device capable of receiving, transmitting, and/or manipulating data for such purposes as, but not limited to, display on a display unit connected thereto.

The World Wide Web has become a popular medium for information exchange. Literally millions of new Web pages have been developed in the past several years as more and more individuals, businesses and organizations have discovered the power of web network. Many of these Web pages are written only in English. Non-English speaking users often have difficulty reading Web pages written in English, and thus may have difficulties to take advantage of information available on the web.

Current automatic translation software which translates text Web pages from a source language such as English to a foreign native language, typically utilize databases that contain information about various languages and a

translation module that refers to this database when performing automatic translation. Utilizing such automatic translation software with Web browser's proxy function enables to translate documents transmitted to the Web browser and display the document translation on the user's screen. Exemplary automatic translation software of this type is "King of Internet Translation Ver 1.x," sold by IBM Japan, Ltd.

Unfortunately, it can be difficult to automatically translate text in one language to text in another language so that the meaning of the original text is accurately reflected in the translation. Further more, it is difficult to phrase correctly the translated text and comply with the grammar rules of the translation language. This may often be a result of the ambiguity inherent in various languages. For example, ambiguity may arise from the use of words that have more then one meaning and that frequently appear in the text to be translated. When translating such word, one must select the appropriate meanings in relation to the sentence context and meaning.

Another source of ambiguity may arise from variations in grammar rule and formats between different languages. English sentences, for example, have specific structural sentence words sequence, such as "subject-verb-object." When pronouns such as "that", "which", and "why" are omitted, understanding English sentence patterns and grammar may be difficult. Words in sentence have different grammar function, and thus must be treated differently. Each word should be analyzed separately and in conjunction with the other words of the sentence in order to attain proper translation.

It is thus a prime object of the invention to avoid at least some of the limitations of the prior art and to provide a method and system for online automatic translation from original language text to any other language.

SUMMARY OF THE INVENTION

A method for translating text sentences from source language to target language using databases including vocabulary and thesaurus of source and target languages, grammar function of each word, translation index, vocabulary of verbs paradigm, vocabulary of preposition, adverb and adjectives inflections, said method comprising the steps of: breaking sentence to text fragments according to punctuation marks; identifying grammar form of text fragments according to verb inflection, punctuation marks and grammar key words; identifying dominant tense form of sentence according to verb inflection and identified grammar form of text fragments; identifying subject of text fragment by locating the word appearing next to the first preposition wherein the exact location of the word (before or after the preposition) is specified according to sentence grammar rules of the source language; locating all verbs in text fragment and translate each verb to source grammar form in target language using translation index; inflecting each translated verb using vocabulary paradigm according to dominant tense form and according to identified subject; locate all nouns in text fragment and translate each noun to source grammar form in target language using translation index; analyzing each noun word grammar form and inflection such as single/plural or male/female; locating all adjectives, prepositions and article words relating to each noun: translating located adjectives, prepositions and article words using translation index according to respective vocabulary and translation index;

inflecting translated adjectives, prepositions and article words according to nouns grammar form using respective vocabulary paradigm; and re-arranging translated words order in each text fragment using grammar rule of target language according to grammar function of each word;

BRIEF DESCRIPTION OF THE DRAWINGS

These and further features and advantages of the invention will become more clearly understood in the light of the ensuing description of a preferred embodiment thereof, given by way of example only, with reference to the accompanying drawings, wherein-

Fig. 1 is a general diagram block of the automatic translation system according to the present invention;

Fig. 2 is a flow-chart illustrating the method of converting web-page text form source language to target language according to the present invention;

Fig 3 is a flow-chart of the sentence translation module according to the present invention;

Fig. 4 is a flow-chart of word translation module according to the present invention;

Fig. 5 is a flow-chart illustrating the method of determining sentence grammar form according to the present invention;

Fig. 6 is a flow-chart illustrating the method of determining dominant tense of text sentence according to the present invention;

Fig. 7 is a flow-chart illustrating the method of determining sentence subject according to the present invention;

Fig. 8 is a flow-chart illustrating the method of rearranging word order in sentence according to the present invention;

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The embodiments of the invention described herein are implemented as logical operations in a computing system. The logical operations of the present invention are presented (1) as a sequence of computer implemented steps running on the computing system and (2) as interconnected machine modules within the computing system. The implementation is a matter of choice dependent on the performance requirements of the computing network system implementing the invention. Accordingly, the logical operations making up the embodiments of the invention described herein are referred to variously as operations, steps, or modules.

Fig.1 block diagram illustrates the structure of web-page translation system. As seen in Fig. 1 conversion module 10 is associated with user browser and controls the operation of the sentence translation module 12 ("Sentence module"). The convector module function is to intercepts all incoming data from network for instance, e-mail, web page etc., detect text data and translate thereof to desired language. (detailed description of the converter module will be described down bellow). The detected text data is analyzed by the sentence module 12 to identify the sentence context and dominant grammar features. The analysis results are used by the word-translating module 14 for selecting and phrasing the proper translation for

each word or idiom. The translating modules 12 and 14 are using different databases containing vocabularies of words for different functions.

Databases 16 and 18 include vocabulary of words of at least two languages wherein key index 26 correlates between corresponding words of any pair of different language. These databases include information of each word grammar function in the sentence such as noun, verbs, adjectives etc. Thus translating modules use these databases not only for translation, but also for detecting the grammar function of the words.

Database or alternatively designated respective modules 20,22,24 and 26 enable to phrase the words in different language according to respective language grammar rules. Database 26 contains vocabulary of idioms for each translated language wherein each idiom contains at least two words.

The translation system according to the present invention can be implemented as software application at the user end, or alternatively as application service at a remote network server such as Internet service provider (ISP).

Fig. 2 illustrates the flow chart of the web page converter. The converter receives any kind of network data such as HTML web-page code, and parses the data to detect text objects designated for screen display. Each text object is examined to determine it's dominant language ("Source language"). The source language is identified according to common words of each language such as "The" or "for" in the English language by using the common word database 24. The converter activates the sentence translation module to translate the text object from the source language to the designated target

language as was pre-defined by the user. The converter module creates new web page based on the original HTML code wherein original text objects are replaced by translated text object as phrased by the Sentence module. Furthermore, alignment and display commands of the HTML code are changed according to target language paragraph format rules.

Fig. 3 illustrates the workflow of the Sentence module. The basic concept of this module is to analyze and parse the text object step by step in order to identify the sentence context and its grammar formats. The order of performing the analysis steps is essential for achieving best translation and phrasing results. The analysis is preformed separately for each sentence part ("Text fragments"), wherein each sentence part is identified by punctuation marks such as ".", "'" etc. Although the translation process is more efficient according to the preferred stages order as suggested according to the present invention, different order of the stages can be used. Moreover, in case of grammar rules of different languages, the order of stages can be changed accordingly.

The first essential stage is determining the dominant sentence grammar format (See step A in Fig. 3) such as imperative, question, passive voice etc. The process of determining said format is illustrated in Fig. 5. The basic parameters used for such analysis are punctuation marks (e.g. "?" or "!"), tense form of verbs and special grammar words such as "be" "was" etc., although the rules for such analysis may be different for each source language the concepts remains the same.

The next stage is to identify the dominant tense form of each text fragment (see step b in Fig. 3). Step B process is illustrated in Fig 6, the

dominant tense form is determined by verb conjugation of all detected verbs and the grammar format as was identified in the first step.

The third essential stage of the process is determining the sentence context, first by identifying the sentence subject (see step C in Fig. 3). The process of step C is illustrated in Fig. 7. The basic idea is to find the dominant word which is the subject of the text fragment. Most frequently the subjected is located after/before the first preposition word in sentence or alternatively after the first verb. The location of the subject is depended on the grammar form of the text fragment, for example if its passive the subject appears after the first verb according to English grammar rules. The rules must be changed according to source language grammar rules. The sentence context can be further determined by key words which are commonly used in specific areas (e.g. computers, medicine etc.).

According to further embodiment of the present invention it is suggested to identify sentence context according to key words given by the author of the web page which are written within the HTML code.

According to furthermore embodiment of the present invention it is suggested to use an idioms database 26 for identifying group of words which have special meanings. Proper translation of said idiom might be essential for identifying the sentence context.

The fourth essential stage of the process is analyzing each of the nouns type and inflection, see step D in Fig. 3. Basically, this process identifies the affixes added (e.g. "s") or alterations of the noun, indicating of plural/ single,

male/female forms. This analysis is essential for the phrasing and inflecting of words relating to the noun such as prepositions, adjectives etc..

Once completing the above analysis, the Sentence module translates each of the text fragment words by activating the word translation module ("Word module"). Fig. 4 illustrates the word translation process. Each word is translated by using the vocabulary database 12, 14 and respective translation index 28. Most frequently, words of the source language has more then one meaning and different synonyms of the words of the target language can be chosen for translation. The preferred translation according to the present invention is determined according to results of the sentence analysis, including sentence context, sentence subject, sentence grammar form, word grammar form and meaning of near by words.

Finally, after all words of the text fragment are translated, the word order must be re-arranged to fit the grammar rules of the target language. This process is illustrated in Fig. 8. The word order in the sentence is determined by the grammar function of each word. In each language there are different rules for word order, hence the location of each word in the sentence must be changed accordingly.

According to further embodiment of the present invention it is suggested to record short sentences original text and respective translation which are frequently translated form one language to another. Maintaining records of such sentences in a designated database can improve the performance of the translating process.

According to another embodiment of the present invention it is suggested to record translation of complete web pages. It is known that some web pages are visited more frequently than other pages. Such pages are usually cached at the end user or alternatively at proxy Internet server (e.g. ISP servers). Therefore it is suggested to store along with the cached web page their respective translation. As a result, time latency of translating web pages is reduced

While the above description contains many specifities, these should not be construed as limitations on the scope of the invention, but rather as exemplifications of the preferred embodiments. Those skilled in the art will envision other possible variations that are within its scope. Accordingly, the scope of the invention should be determined not by the embodiment illustrated, but by the appended claims and their legal equivalents.